VIDEO KEY FRAME SELECTION BY CLUSTERING WAVELET COEFFICIENTS

Satoshi Hasebe, Makoto Nagumo, Shogo Muramatsu and Hisakazu Kikuchi

Department of Electrical and Electronic Engineering
Niigata University
Niigata 950-2181, JAPAN
Phone: +81 25 264 2206, Fax: +81 25 264 2130
hasebe@telecom0.eng.niigata-u.ac.jp

ABSTRACT

This paper presents a new method for selecting key frames from a given video sequence. It is characterized by the fact that the entire process works in a wavelet transform domain. At first, shot boundaries are detected to define initial key frames. Secondly, specified but any number of key frames are selected by clustering feature vectors. Its effectiveness is evaluated in terms of precision rates and processing speeds. The proposed method offers more satisfactory results and works faster than the other existing methods.

1. INTRODUCTION

Key frames are the most common representation for the abstraction of video sequences. They help a user understand the video contents quickly by tiling key frames on a display[1][2]. Key frame production is the fundamentals for video content analysis[3] and querying techniques[4][5] that use frame features such as colors, shapes and textures.

Various kinds of techniques for selecting key frames have been proposed. Most of them selects initial key frames, and then reduce the number of key frames by clustering feature vectors. Typical strategies for selecting initial key frames are:

(a) Select all frames in a video sequence.

(b) Select regularly subsampled frames.

(c) Select a set of initial key frames according to the result of shot boundary detection.

Method (a) involves a large number of frames in clustering. Thus, the dimension of the feature vector should be adequately small to complete clustering in practical time. Such a method is not suited for on-line video analysis[6], where processing speed is crucial. Method (b) reduces the number of frames simply. Redundant frames are removed by subsampling to some extent, since neighboring frames are very similar to each other. It can still leave redundant frames, and can falsely remove important frames, because it does not consider the frame contents at all. Method (c) selects initial key frames by considering the frame contents at the expense of computational cost. The proposed method is based on Method (c). It successfully reduces the computational cost by making efficient feature vectors in shot boundary detection, and they are kept to be used in clustering.

Lee et al.[7] uses RGB histograms as high dimensional feature vectors that are extracted from initial key frames. Those initial key frames are selected according to the result of shot boundary detection. Singular value decomposition is applied to those feature vectors before k-means clustering is performed. This contributes to speed up the clustering.

Drew et al.[8] defines *chromaticity signatures* of initial key frames by a few bases to form low dimensional feature vectors. Those bases are prepared in advance according to the result of the singular value decomposition. After adjacent clusters are merged, non adjacent clusters are merged. Finally, those frames of which feature vectors are closest to the cluster centers are selected as key frames. Both algorithms require full-decoded frame pictures for extracting feature vectors.

The purpose of this study is to select key frames from a given video sequence. First of all, initial key frames are selected according to the result of shot boundary detection. Secondly, any number of key frames are selected by clustering feature vectors of those initial key frames, and the number can be as many as desired. Our contribution in this work is to conduct both shot boundary detection and clustering in a wavelet transform domain.

A video sequence is assumed to be encoded by wavelet-based coding techniques such as Motion-JPEG2000[9]. JPEG2000[10] is the next generation image and video coding standard. It offers a variety of scalabilities. Two-dimensional wavelet coefficients are obtained by partial-decoding of JPEG2000 bit streams. In spite of the transform domain processing, no inverse wavelet transform is needed, and this fact leads to less computational cost and memories.

2. KEY FRAME SELECTION

2.1 Initial Key Frames

When a video sequence is given, shot boundaries are to be detected to find initial key frames. We employ a two-step shot boundary detection algorithm[11], which works in a wavelet transform domain. It captures gradual shot transitions as well as abrupt shot transitions. It computes a distance between video intervals to find a isolated interval. Then, it computes a distance between frames to specify an exact location of a shot boundary.

The two-step shot boundary detection algorithm takes a single threshold value to control the sensitivity of shot boundary detection. The threshold value should be as low as possible to avoid detection misses. Although a lower threshold value may cause false positives, redundant frames are removed in the process of clustering. So it hardly influence the performance unless an extremely huge number of false positives are produced.

After a given video sequence is divided into multiple shots, a single frame that well represents the shot content is selected for every shot. A frame adjacent to a shot boundary is likely to belong to a shot transition. Indeed, gradual transitions such as dissolves and fades consists of several frames. Such a frame that involves different two shot contents is not qualified for a representative frame. Hence we select an initial key frame that is located at the midpoint between one shot boundary and the next shot boundary.

2.2 Feature Vectors of Key Frames

Feature vectors of initial key frames are constructed in the process of shot boundary detection, where they are used to compute distance between frames to find significant changes. We intend to perform clustering with those feature vectors. For this purpose, similarity distance and average feature vector are defined.

A feature vector F comprises the coarsest subband C and the *significance map*[11] S of finer subbands of the two-dimensional wavelet transform of a frame picture.

$$F = \{C, S\}. \tag{1}$$

The coarsest subband consists of quantized wavelet coefficients. It is a large scale approximation of a frame. The significance map is a binary map: significant coefficients in finer subbands are encoded as unity and insignificant coefficients are encoded as zero. It implies the presence of sharp changes such as edges and textures.

The distance between two feature vectors, F_m and F_n, is defined after some preliminary definitions. The L1 distance between two coarsest subbands, C_m and C_n, is described by

$$||C_m - C_n||_{L1} = \sum_i \sum_j |c_m(i,j) - c_n(i,j)|, \tag{2}$$

where $c(i,j)$ denotes a quantized coefficient at (i,j) in the coarsest subband. The Hamming distance between two significance maps, S_m and S_n, is given by

$$||S_m - S_n||_H = \sum_i \sum_j \{s_m(i,j) \oplus s_n(i,j)\}, \tag{3}$$

where $s(i,j)$ denotes a binary at (i,j), and $\oplus$ represents exclusive OR. Finally, the distance between two feature vectors, F_m and F_n, is defined by the weighted sum of Eq. (2) and Eq. (3), as follows.

$$||F_m - F_n|| = w_0||C_m - C_n||_{L1} + w_1||S_m - S_n||_H, \tag{4}$$

where w_0 and w_1 are weights.

A cluster center is defined by the average of feature vectors in the cluster. Given multiple feature vectors, $F_1, F_2, \cdots, F_n$, average vector $\bar{F}$ is defined as follows. Average coarsest subband $\bar{C}$ of $C_1, C_2, \cdots, C_n$ is calculated by

$$\bar{c}(i,j) = \frac{1}{n}\sum_{k=1}^{n} c_k(i,j), \tag{5}$$

where $\bar{c}(i,j)$ denotes a coefficient at (i,j) in $\bar{C}$ and $c_k(i,j)$ denotes a coefficient at (i,j) in C_k. Average significance map $\bar{S}$ of $S_1, S_2, \cdots, S_n$ is calculated by

$$\bar{S} = T(H), \tag{6}$$

$$h(i,j) = \sum_{k=1}^{n} s_k(i,j), \tag{7}$$

where $h(i,j)$ denotes the number of significant coefficients that are located at (i,j), and $s_k(i,j)$ is a binary at (i,j) in S_k. A mapping $T(\cdot)$ shows a thresholding-after-sorting operation as follows:

> After every element h(i, j) in the argument H are sorted by their values in descending order, the largest N elements are quantized into one and the others are quantized into zero. N is a given constant.

The resulting average coarsest subband and the significance map compose the average feature vector.

$$\bar{F} = \{\bar{C}, \bar{S}\}. \tag{8}$$

2.3 Clustering of Feature Vectors

In the *k*-means clustering algorithm, both the result of clustering and the number of iterations depend on the initial cluster centers. It is desirable that the initial cluster centers spread over the feature space as widely as possible. Since adjacent frames in a video sequence are similar to each other, they are also likely to be located closely in the feature space. Hence we regularly subsample the sequence of feature vectors of initial key frames so that a specified number of feature vectors are obtained. Those feature vectors are treated as the initial cluster centers.

For a given set of feature vectors of initial key frames and initial cluster centers, the *k*-means algorithm is performed as follows.

STEP 1 For every feature vector, compute the distance between each cluster center and the feature vector by Eq. (4). Then, associate every feature vector with the closest cluster.

STEP 2 For every cluster, compute a new cluster center according to Eq. (6) and Eq. (7).

STEP 3 For every cluster, compute the distance between the current cluster center and the previous cluster center by Eq. (4). If all of the distances are small enough, proceed to STEP 4. Otherwise, go back to STEP 2.

STEP 4 For every cluster, find the feature vector that is closest to the cluster center. This feature vector is the representative feature vector of the cluster.

As a result of the four-step procedure, any desired number of feature vectors are obtained. Those frames corresponding to the obtained feature vectors are the final key frames.

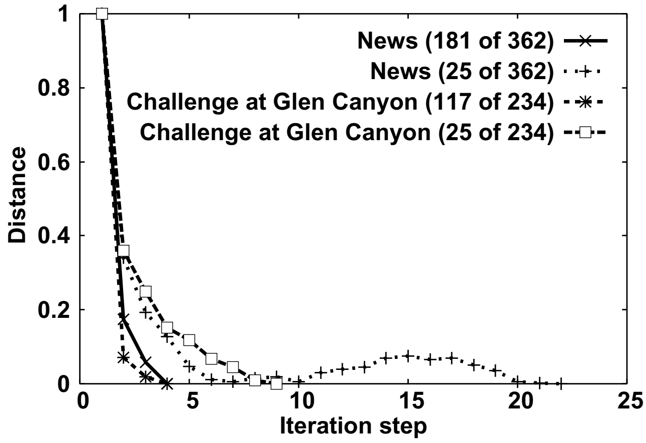

Figure 1: The Average of Distances Between Previous and Current Cluster Centers Versus the Iteration Step.

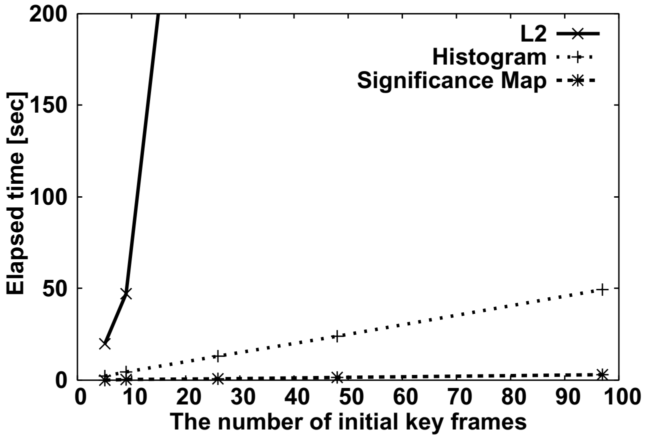

Figure 2: Elapsed Time Versus the Number of Initial Key Frames.

3. EXPERIMENTAL RESULTS

As an experiment, we have tried to pick up a fixed number of key frames. It is 25. The other key frame selection experiment is to select half the number of initial key frames for a test video sequence. Those initial key frames are selected according to the result of shot boundary detection, and their total number is equal to that of detected shots.

Figure 1 shows the average of distances between previous and current cluster centers at each iteration step. Every distance is normalized so that the maximum distance takes the value of one. *News* has 362 initial key frames. They are reduced into 182 and 25 frames, respectively. Similarly, *Challenge at Glen Canyon* has 234 initial key frames to be reduced into 117 and 25 frames, respectively. The algorithm has been iterated until the distance reaches zero. Smaller destination number of key frames requires more iterations. As seen in Fig.1, at most 10 iterations are sufficient.

If just a single key frame has been selected in a shot, and if it belongs to a stable non-transient interval, it is considered as a valid key frame. If two or more key frames are selected from a single shot, the first key frame is considered as valid, and the other key frames are considered as invalid. The total number of valid key frames is equal to the number of shots in a test video sequence. To evaluate the validity of selected key frames, recall and precision rates are calculated as follows. Recall is defined by the percentage of valid key frames that are actually selected among all valid key frames. Precision is defined by the percentage of valid key frames among all of key frames that are actually selected.

Table 1 lists the recall and precision rates of initial key frame selection for three test sequences. Because of detection misses and false positives, neigher recall nor precision rates reach 100%. *Color Harmony for Your Home* has low recall rates because the shot boundary detection is performed with a relatively high threshold value.

Since the number of key frames is reduced as the result of clustering, recall rates naturally drops. On the other hand, owing to the reduction of invalid key frames by clustering, precision rates will improve. We are not interested in the recall rates of selected key frames. Table 2 lists the precision rates of key frame selection.

Two pixel domain methods and the proposed wavelet domain method are compared. A method based on the L2 distance uses pixels themselves as the feature vector, and the L2 distance as the distance. Histogram-based method uses RGB histograms as the feature vector, and the absolute sum of difference between corresponding histogram bins as the distance. The proposed method uses the feature vector defined by Eq. (1) and the distance defined by Eq. (4). All those methods employ *k*-means as the clustering algorithm. The method based on L2 distance outperforms the others in our experiments.

Figure 2 shows the elapsed time in clustering versus the number of initial key frames to be processed. Shot boundary detection is performed with several different threshold values to obtain initial key frames, and the number of key frames are halved by clustering. It should be noted that the time needed for the shot boundary detection is not included in the plots.

The method based on L2 distance is not suited for clustering many key frames owing to the huge cost of a distance calculation. The fact that the size of a feature vector is too large to process on memory worsen the performance. The processing time of the other two method grows slowly. Especially, the proposed method based on the significance map is outstandingly fast. This is because the feature vectors are obtained in the process of shot boundary detection.

4. CONCLUDING REMARKS

We have proposed a new method for selecting key frames from a given video sequence. It works entirely in a wavelet transform domain. Shot boundary detection is a preprocessing, and is followed by the *k*-means clustering of feature vectors of key frame candidates to find key frames.

Precision rates demonstrate that the proposed method offers satisfactory results compared with other methods. Moreover, the proposed method dose not require inverse wavelet

Table 1: Recall and Precision Rates of Initial Key Frame Selection.

Video Sequence	The Number of Initial Key Frames	Recall	Precision
Challenge at Glen Canyon	234	93.8%	97.0%
Color Harmony for Your Home	93	40.0%	83.9%
News	362	91.8%	86.2%

Table 2: Precision Rates of Key Frame Selection.

(a) Results of Selecting Half the Number of Initial Key Frames

Video Sequence	L2 Distance	Histogram	Proposed
Challenge at Glen Canyon	98.3%	98.3%	97.4%
Color Harmony for Your Home	95.7%	91.3%	91.3%
News	93.4%	93.4%	92.3%

(b) Results of Selecting 25 Key Frames

Video Sequence	L2 Distance	Histogram	Proposed
Challenge at Glen Canyon	100.0%	96.0%	100.0%
Color Harmony for Your Home	96.0%	96.0%	92.0%
News	100.0%	96.0%	96.0%

transforms, and the feature vectors can have been generated in the process of shot boundary detection. A significant reduction in computational cost and memory requirements is gained by this combination of the transform-domain shot boundary detection and key frame selection.

REFERENCES

[1] S. Uchihashi, J. Foote, A. Girgensohn, and J.Boreczky, “Video Manga: Generating Semantically Meaningful Video Summaries”, *Proc. of the Seventh ACM International Conference on Multimedia*, pp.383–392, Orlando, 1999

[2] J. Boreczky, A. Girgensohn, G. Golovchinsky and S. Uchihashi, “An Interactive Comic Book Presentation for Exploring Video”, *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp.185–192, The Hague, 2000

[3] L. Chaisorn, T. -S. Chua, and C.-H. Lee, “Segmenting Stories in News Video”, *Handbook of Video Databases Design and Applications*, pp.1133–1148, CRC Press, 2004

[4] S. Hasebe, S. Muramatsu, S. Sasaki, H. Kikuchi, “Video Querying Based on Three-Dimensional Wavelet Transforms”, *Proc. of ITC/CSCC2001*, pp.1196–1199, 2001.

[5] X. Wen, T. D. Huffmire, H. H. Hu, A. Finkelstein, “Wavelet-Based Video Indexing and Querying”, Multimedia Systems, 7, 5, pp.350-358, Springer-Verlag Heidelberg, 1999.

[6] W. Zhou and C. -C. J. Kuo, *Intelligent Systems for Video Analisys and Access Over the Internet*, Prentice Hall, 2002

[7] S. Lee and M. H. Hayes, “A Fast Clustering Algorithm for Video Abstraction” *Proc. IEEE ICIP*, Barcelona, 2003

[8] M. S. Drew and J. Au, “Video Keyframe Production by Efficient Clustering of Compressed Chromaticity Signatures” *Proc. of the Eighth ACM International Conference on Multimedia*, pp.365-367, Marina del Rey, 2000

[9] T. Fukuhara, K. Katoh, S. Kimura, K. Hosaka and A. Leung, “Motion-JPEG2000 Standardization and Target Market”, *Proc. IEEE ICIP*, No.TA0208, Vancouver, 2000.

[10] D. S. Taubman and M. W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, Massachusetts, 2002.

[11] S. Hasebe, S. Muramatsu, S. Sasaki, J. Zhou and H. Kikuchi, “Two-Step Algorithm for Detecting Video Shot Boundaries in a Wavelet Transform Domain”, *Proc. 3rd International Symposium on Image and Signal Processing and Analysis (ISPA03)*, pp.245–250, Rome, 2003.